http://dx.doi.org/10.5392/IJoC.2015.11.2.015

Local Similarity based Document Layout Analysis using Improved ARLSA

Gwangbok Kim, SooHyung Kim, InSeop Na*
School of Electronics and Computer Engineering
Chonnam National University, Gwangju 500-757, Korea

ABSTRACT

In this paper, we propose an efficient document layout analysis algorithm that includes table detection. Typical methods of document layout analysis use the height and gap between words or columns. To correspond to the various styles and sizes of documents, we propose an algorithm that uses the mean value of the distance transform representing thickness and compare with components in the local area. With this algorithm, we combine a table detection algorithm using the same feature as that of the text classifier. Table candidates, separators, and big components are isolated from the image using Connected Component Analysis (CCA) and distance transform. The key idea of text classification is that the characteristics of the text parallel components that have a similar thickness and height. In order to estimate local similarity, we detect a text region using an adaptive searching window size. An improved adaptive run-length smoothing algorithm (ARLSA) was proposed to create the proper boundary of a text zone and non-text zone. Results from experiments on the ICDAR2009 page segmentation competition test set and our dataset demonstrate the superiority of our dataset through f-measure comparison with other algorithms.

***Key words**: Document Layout Analysis, Page Segmentation, Table Detection, Adaptive RLSA.*

1. INTRODUCTION

Document Analysis is to recognize text and group arranged in order of reading. When we convert printed document into digital file obtained by scanner or smartphone camera, first step is analyzing component and properties of document layout. There are many studies on this subject. The bottom-up approach is start at small objects, such as pixels, zones. After that, objects having same properties and adjacent each other are combined together. Another approach, top-down algorithm has begun from the divided regions of whole image or high-level region. And each block or region is classified into each property such as text, graph, table, separator image and noise.

The bottom-up approach are both the oldest [1], [2] and more recently published [3]. They classify components from small zone and make group with them. The Key advantage of bottom-up methods is that they can handle arbitrarily shaped regions with ease [4]. Top-down approach starts at largest structures on document, and determine properties of each group or region. However, these methods are unstable for using variety of formats. Recently, hybrid approach overcomes many problems of this research field. The whitespace can be used for detecting group and find the polygon boundary for each region [4]. In paper [5], they cluster the text candidates using mean-shift analysis technique according to their corresponding sizes. The main key of this paper is global stroke width variance features. The efficient morphological-based method is employed to eliminate text and line candidates. An author [6] proposed a method using efficient white space. He segments whitespace rectangles using connected component analysis. After that, the whitespace rectangles are filtered by foreground and background information such that the remaining rectangles are likely to form column separators.

In general, layout analysis algorithms deal with segment text zone and non-text. In case of research for detecting table, almost paper detached the table detection part from layout analysis system. Faisal Shafait [7] takes a practical algorithm for table detection. The author built upon two components of the layout analysis module, column partitions and column layout. They consider various types of documents and used page column split and join it together. In paper [8], they proposed table detection in noisy off-line handwritten documents. Following a bottom-up approach, they divide the page into small tiles. After that, structural features such as Gradient-Structural-Concavity are used for SVM training.

This paper is organized as follows. In Section 2, we describe a proposed method for layout analysis and table detection in detail. And we provide experimental result of our implemented algorithm in Section 3, and finally Section 4 describes the conclusions.

* *Corresponding author, Email: ypencil@daum.net*
Manuscript received Mar. 16, 2015; revised May. 29, 2015; accepted Jun. 10, 2015

2. PROPOSED METHOD

The proposed layout analysis is based on distance transform. Fig. 1 shows flowchart of our whole system. At first, we convert input image, color image to binary image. Connected component analysis (CCA) with distance transform algorithm is used for detecting separator images and table candidates. The components except table candidate and separator candidates are classified by window having adaptive size. After that, text components should be merged using smoothing approach called ARLSA. And, non-text image is obtained by difference image between original image pixels and text pixels. Finally, we detect table zone using non-text pixel ratio and text pixel ratio.

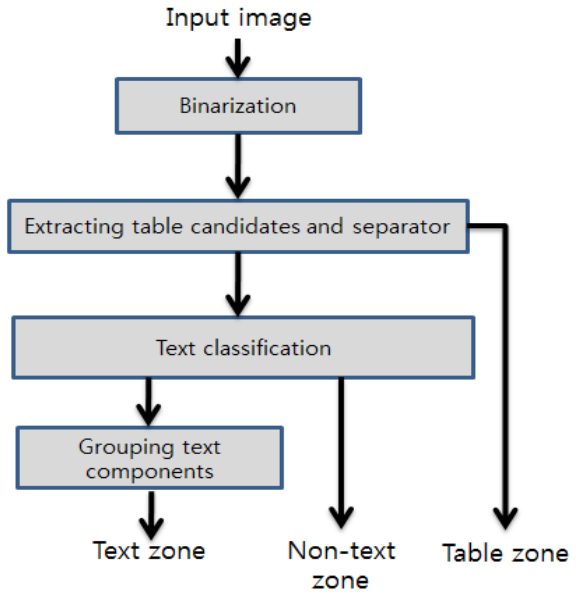


Fig. 1. Flowchart of the proposed algorithm

2.1 Extraction of table and separator

The Connected Component Analysis (CCA) is an approach to generate groups consisting of adjacent pixels and utilize them. Before main step, we separate line separator components and big graphic regions from others. It is important to detect text components and make blobs. The connected components (CCs) are filtered by maximum length of height or width. The line separator components are classified easily by aspect ratio. Almost all of separator components has high aspect ratio as about 8~10.

The table detection is based on mean value of distance transform. The paper [5] also employs global distance variance to determine text region. In our paper, mean value of distance transform for each connected component shows high similarity. At the first step, we extract components having low distance value, height and width on mean distance value. In order to detect exact table region, we employed decision tree using heuristic predefine parameter and achieved good result on our dataset. Fig. 2(b) shows the pixels extracted by predefine threshold. It is no matter how the threshold value is accurate since post processing detects the table region from candidates. As shown in Fig. 2(d), we detect table regions using text ratio and non-text ratio. After finding out text component and non-text component, rectangle area of each candidate is estimated using text ratio and non-text ratio.

The table has characteristics based on rectangle shape as follows. Low pixel ratio of text components within rectangle: In table region, text pixel appears less than regular value. Because, there are many white space and tab-stops outside of each table cell regardless of frame components.

1) There is no non-text pixel within rectangle: Most tables don't include non-text components. However, we consider a miserly pixel ratio value of non-txt due to miss-detection of non-text.
2) Low pixel ratio of frame: table has thin frame within table region. The table has various types such as an image including only vertical lines or only horizontal line. Also, there is a case of image including no line. They are generated from binarization method or design make it originally. Basically, we extract candidates of table from distance transform image and CCA previously.

After that, those components adjacent or touched with separator would be joined together since first line of table appeared like separator by result of binarization algorithm. Therefore, we combine table zone and adjacent separators.

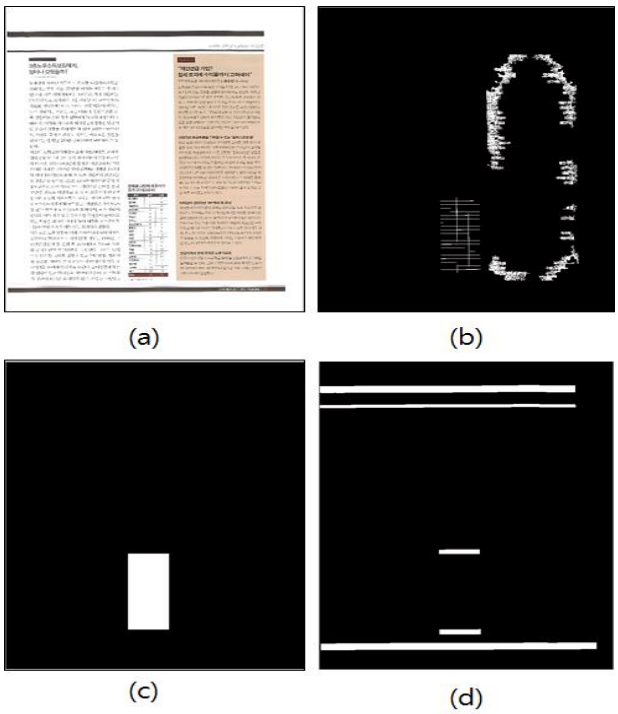


Fig. 2. (a) Original Image, (b) Candidates of table filtered by CCA, (c) table region result, (d) Separator Components

2.2 Classification of text and non-text

Texts are parallel each other and have similar size and thickness. As discuss in Section 2.1, we classify the text component using mean distance transform image and height of each component. However, stroke width information solve this problem and distinct text and non-text using the distance transform. The distance transform represents how far each pixel

from boundary. To utilize it, we blurred distance transform image and get a mean value of local area.

Some components are touched originally or generated by low resolution of image. The predefine limitation of window size for searching neighboring CCs makes the problem that there are no neighboring CCs within window. As shown in Fig. 3, a datum component for estimating region consists of two letters 't' and 'h'. It is the result from binary-zation method and connected component analysis. However, our proposed method searches the adaptive range corresponding to size of datum component. The proposed adaptive window grows repeatedly until neighbor component belongs. If some reference points of others exist inside the searching window, they are used for estimating local similarity. The text components keep the similar line on middle of window as Fig. 3. And, these have similar height with neighbor components. Basically, the window size and the number of components are determined by heuristic condition. However, we resize the window until 2 neighbor components are with range when it doesn't meet first condition. Additionally, we ignore the noise and component dot symbol such as a component of alphabet 'i' or a comma.

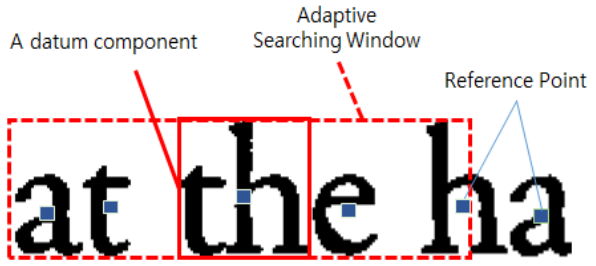


Fig. 3. Examples of adaptive searching window

2.3 Text grouping by an improved ARLSA

We improve the adaptive run length smoothing algorithm [10] to make text lines on an image. It makes better result and stable smoothing for next step. The RLSA [9] is one of the most common algorithms used in page layout analysis and segmentation techniques. The RLSA algorithm links a pixel to a pixel by sequence of background pixels when the distance between two parallel black pixels is less than threshold value. However, it is not sensitive for various font size and image size as Fig. 3. We can see the image box touched with text lines in Fig. 4.

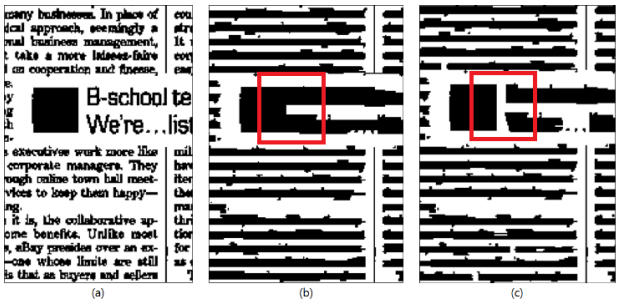


Fig. 4. (a) Binary Image, (b) Smoothing by ARLSA, (c) Smoothing by proposed ARLSA

The adaptive run length smoothing algorithm (ARLSA) aim is to overcome several problems. They considered height ratio between two black pixels to connect after CCA.

Paper [10] considered for following metrics:

L(S) : length of the sequence, that is the number of white pixels

H(S) : height ratio between two CCs

O(S): the horizontal overlapping between the bounding boxes of two CCs.

N(S): If the 3×3 neighborhood of at least on pixel of the sequence S(i, j), N(S) is set to 0. If there is third connected component, N(S) is set to 1.

In our paper, we used L(S), H(S) and O(S) of exist algorithm [10]. The height ratio and the overlapping height between two CCs are good feature to determine whether or not two components are merged. We modified the constraint O(S) to connect two components by vertical direction and except N(S) metrics. Fig. 4 shows a problem of ARLSA and result of our improved ARLSA. The box graphic on image is touched with text lines despite two components are not same properties. In paper [10], they predefine parameter for each metrics by experiment such as T_h (height ratio) as 3 between two components. It makes a problem like Fig. 4(b). Therefore, we propose an additional metric D(S) to connect two similar components each other. The distance transform algorithm is good feature to compare with neighbor components. The metric based on distance transform, D(S) is a similar feature with stroke width value of connected component analysis. Therefore, we can solve the problem using metric D(S). Algorithm 1 shows how our ARLSA makes text grouping.

Algorithm 1 – Adaptive Run Length Smoothing Algorithm

```
1:  if current_pixel == black_pixel
2:     tempX = current_pixel.x
3:     tempY = current_pixel.y
4:     tempRect = ConnectComponent(currentPixel)
5:     tempD = current_pixel.distance_transform
6:     for left to right
7:        if current_pixel == black_pixel
8:           L(S) = tempX – current_pixel.x
9:           H(S) = tempY/current_pixel.y
10:          O(S) = calculation(tempRect, currentRect)
11:          D(S) = tempD/current_pixel.distance_transform
12:           if L(S) < T_l & H(S) < T_h & O(S) < T_o &
      D(S) < T_d
13:              connect(temp_pixel, current_pixel);
14:           end
15:       end
16:    end
17: end
```

3. EXPERIMENTAL RESULT

We evaluated the document layout analysis performance of the proposed method on the ICDAR2009 page segmentation competition dataset. The images are in full color and similar size, 2327×3132 averagely. Our system is coded with Matlab

2012 in Windows on a PC with Intel(R) Core(TM) 2 CPU-3.0GHz. Table detection performance achieves high score in our dataset. For 18 images of our dataset including at least a table, 94% is success in detection shown as Fig. 5.

To evaluate performance of our page segmentation algorithm, the F-measure is employed. The ICDAR2009 Competition also shows F-measure result of all algorithms. Therefore, we use this data to compare with our proposed algorithm. Fig. 6 shows that our result provides the best performance among others by 95.53 for text region and 79.34 for non-text region. And, Fig. 7 shows stable segmentation result using red and blue color boundary. Furthermore, we analyze our algorithm performance and processing time when we reduce an input image size for same picture shown as Table 1. Low resolution images also make a good result. A quarter of image size also makes a not bad result and reduces processing time a lot.

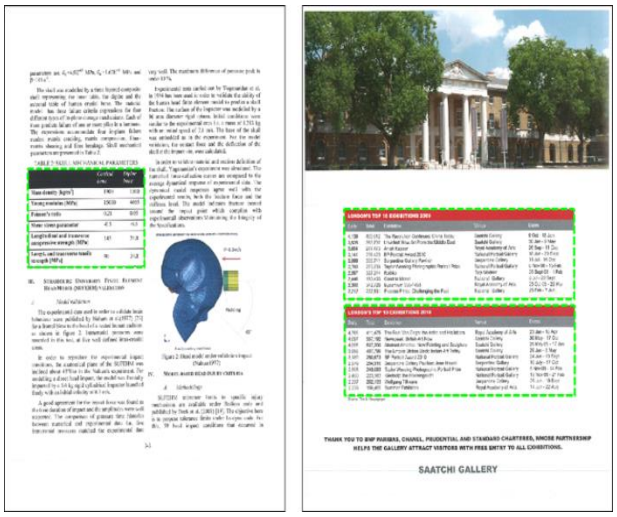
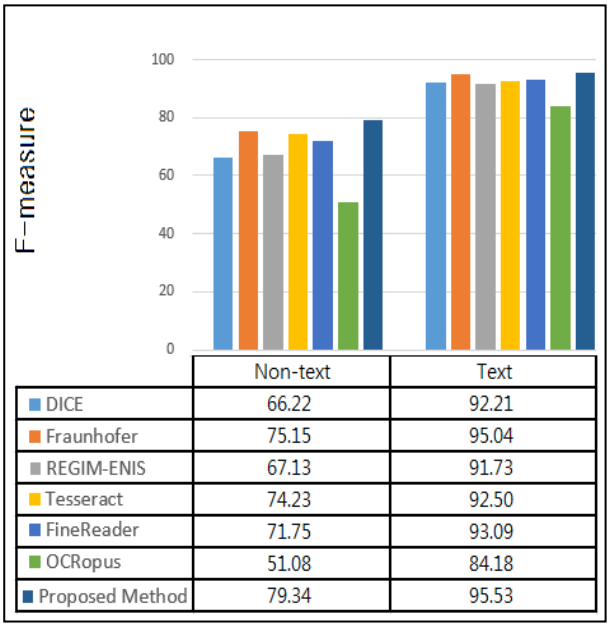


Fig. 5. Examples of table detection for our dataset.



	Non-text	Text
DICE	66.22	92.21
Fraunhofer	75.15	95.04
REGIM-ENIS	67.13	91.73
Tesseract	74.23	92.50
FineReader	71.75	93.09
OCRopus	51.08	84.18
Proposed Method	79.34	95.53

Fig. 6. F-measure evaluation with other methods.

4. CONCLUSION

We presented adaptive window based text detection method and the distance transform based table detection. Furthermore, our algorithm works well on low resolution image and it is fast to run on mobile device. Experimental results on ICDAR 2009 dataset and our dataset have demonstrated the speed and superiority of our algorithm. In the future, we will improve grouping algorithm for noise, separator and reading ordering.

Table 1. F-measure and processing time for various resolutions

Image size	Average F-measure for text region	Average Processing Time(sec)
1/2 (1575 X 1163)	95.5	33.8
1/3 (1050 X 775)	94.0	10.1
1/4 (788 X 582)	92.6	4.3

Fig. 7. Examples of page segmentation result

ACKNOWLEDGEMENT

"This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2015-018993)"

REFERENCES

[1] M. Chen, X. Ding, and Y. Wu, "Unified HMM-based Layout Analysis Framework and Algorithm," Science in China Series F: Information Sciences, vol. 46, no. 6, Dec. 2003, pp. 401-408.

[2] S. P. Chowdhury, S. Mandal, A. K. Das, and B. Chanada, "Segmentation of Text and Graphics from Document Images," ICDAR 2007, pp. 619-623.

[3] A. M. Vil'kin and I.V. Safonov, "Bottom-up Document Segmentation Method based on Textural Features," Pattern Recognition and Image Analysis, vol. 21, no. 3, Sep. 2011, pp.565-568.

[4] R. Smith, "Hybrid Page Layout Analysis via Tab-Stop Detection," ICDAR'09. 10th International conference on. 2009, pp. 241-245.

[5] M. Felhi, S. Tabbone, and M. V. O. Segovia, "Multiscale Stroke-based Page Segmentation Approach," In Document Analysis Systems (DAS), 2014, pp. 6-10.

[6] K. Chen, F. Yin, and C. L. Liu, "Hybrid Page Segmentation with Efficient Whitespace Rectangles Extraction and Grouping," In Document Analysis and Recognition (ICDAR), 2013, pp. 958-962.

[7] F. Shafait and R. Smith, "Table detection in heterogeneous document," Proc. 9th IAPR International Workshop on Document Analysis Systems, 2010, pp. 65-72.

[8] J. Chen and D. Lopresti, "Table Detection in Noisy Off-line Handwritten Documents," In Document Analysis and Recognition (ICDAR), International Conference on. 2010, pp. 399-403.

[9] F. M. Wahl, K.Y. Wong, and R. G. Casey, "Block Segmentation and Text Extraction in Mixed Text/Image Documents," Computer graphics and image processing, vol. 20, no. 4, Dec. 1982, pp. 375-390.

[10] N. Nikolaou, M. Makridis, B. Gatos, N. Stamatopoulos, and N. Papamarkos, "Segmentation of Historical Machine-printed Documents using Adaptive Run Length Smoothing and Skeleton Segmentation Paths," Image and Vision Computing, vol. 28, no. 4, Apr. 2010, pp. 590-604.

GwangBok Kim

He received the B.S., M.S in computer science from Chonnam National University, Korea in 2013, 2015, respectively. He is currently a Ph. D student at Dept. of Electronics and computer Engineering, Chonnam National university, Korea. His main research interests include image processing, pattern recognition and text recognition.

SooHyung Kim

He received his B.S. degree in Computer Engineering from Seoul Nati onal University in 1986, and his M.S. and Ph.D degrees in Computer Science from Korea Advanced Institute of Science and Technology in 1988 and 1993, respectively. From 1990 to 1996, he was a senior member of research staff in Multimedia Research Center of Samsung Electronics Co., Korea. Since 1997, he has been a professor in the Department of Computer Science, Chonnam National University, Korea. His research interests are pattern recognition, image processing, computer vision, and machine learning.

InSeop Na

He received his B.S., M.S. and Ph.D. degree in Computer Science from Chonnam National University, Korea in 1997, 1999 and 2008, respectively. Since 2012, he has been a research professor in Department of Computer Science, Chonnam National University, Korea. His research interests are image processing, pattern recognition, and machine learning.